DOCUMENT RESUME

ED 344 936 TM 018 298

AUTHOR Ackerman, Terry A.
TITLE Assessing Construct Validity Using Multidimensional Item Response Theory.
PUB DATE Apr 92
NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Construct Validity; Equations (Mathematics); *Estimation (Mathematics); Item Bias; *Item Response Theory; *Mathematical Models; Multidimensional Scaling; Raw Scores; Test Construction; *Test Items; *Vectors (Mathematics)
IDENTIFIERS Item Parameters; *Validity Sector

ABSTRACT

The concept of a user-specified validity sector is discussed. The idea of the validity sector combines the work of M. D. Reckase (1986) and R. Shealy and W. Stout (1991). Reckase developed a methodology to represent an item in a multidimensional latent space as a vector. Item vectors are computed using multidimensional item response theory item parameter estimates obtained from calibration programs such as NOHARM II (C. Fraser, 1983). The direction indicated by an item's vector indicates the composite of skills that the item is sensitive to measuring. Shealy and Stout developed a procedure (Simultaneous Item Bias) to detect item bias that encourages the user to identify (and condition on the scores from) only the most valid test items. T. A. Ackerman (1991) suggested that the most valid items for a given test could easily be identified from a plot of the item vectors. That is, the construct valid items should be measuring similar composites and thus lie within a definable sector. Items that lie outside this sector are assessing unintended-to-be-measured skills, and this could be considered construct invalid. This paper describes several uses of the validity sector, including how it can be used to construct tests, detect biased items, and define the raw score scale for a test. One table and 14 vector plots support the discussion. There is a 15-item list of references. (Author/SLD)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

TERRY A. ACKERMAN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED344936

# Assessing Construct Validity using Multidimensional Item Response Theory

Terry A. Ackerman
University of Illinois

Paper presented at the 1992 American Educational Research Association Annual Meeting, San Francisco, CA., April 21, 1992.

TM018298

BEST COPY AVAILABLE

**Abstract**

In this paper the concept of a user specified *validity sector* is discussed. The idea of the validity sector combines the work of Reckase (1986) and Shealy and Stout (1991). Reckase developed methodology to represent an item in a multidimensional latent space as a vector. Item vectors are computed using multidimensional item response theory item parameter estimates obtained from such calibration programs as NOHARM II(Fraser, 1983). The direction indicated by an item's vector indicates the composite of skills that the item is sensitive to measuring. Shealy and Stout developed a procedure to detect test bias (SIB) which encourages the user to identify (and condition on the scores from) only the most valid items. Ackerman (1991) suggested that the most valid items for a given test could be easily identified from a plot of the item vectors. That is, the construct valid items should be measuring similar composites and thus lie within a definable sector. Items which lie outside this sector are assessing unintended-to-be-measured skills, and thus could be considered to be construct invalid. This paper describes several uses of the validity sector including how it can be used to construct tests, to detect biased items, and to define the raw score scale for a test.

# Assessing Construct Validity using Multidimensional Item Response Theory

In their item bias detection procedure Shealy and Stout (1991) call upon the practitioner to identify the most "valid" items to compute a single score for matching subjects. While testing practitioners would like to believe that all of their items are measuring only one valid skill, statistical analyses may suggest otherwise. If one or more items are capable of measuring multiple skills the potential for bias exists. Item bias occurs when items discriminate between levels of a nuisance ability for which two groups of interest have different underlying ability distributions. By conditioning on only the most valid items the procedure is creating a more unidimensional matching criterion and thus increasing the sensitivity to any differential group performance on items that are influenced significantly by nuisance skills. Ackerman (1992) merged the concept of Shealy and Stout's valid test items with the graphic representation of multidimensional items developed by Reckase (1986). Reckase developed the methodology to represent the composite of skills an item is measuring in terms of a vector in a specified latent space. By examining the spread of the item vectors, researchers can gain insight regarding the multidimensionality of a test. If the results of a test are reported as a single score, the user is implicitly assuming that items which contribute to that score are essentially unidimensional. Thus, graphically, the item vectors should lie within a narrow sector called the ***validity sector***. Items lying within the sector are considered to be construct valid as they only distinguish between levels of the intended-to-be-measured traits. Items lying outside of this sector are labeled construct invalid because of their propensity to also discriminate between subjects' abilities on nonvalid or nuisance skills.

The purpose of this paper is twofold. First, I wish to encourage a mutual effort by item writers and psychometricians to achieve a better understanding of what tests are actually measuring. Ideally, there should be a continuing dialogue between the two groups so that content and construct guidelines, followed by item writers, can be statistically verified by psychometricians. The second objective is to provide examples of different ways the concept of the validity sector can be used to help bridge the substantive-statistical gap. Two possible uses

will be discussed: statistically confirming the table of content specifications used in test construction and identifying suspect items for item bias studies.

## Theoretical Background

*Multidimensional Item Response Theory*

The first step in determining the validity sector for a test or subtest is to identify the dimensionality of the latent space. Unfortunately, as anyone who has worked with real data will confess, datasets composed of only ones and zeroes do not have any clear identifying patterns that delineate their dimensionality. It would be nice to scan the data as if it were one giant UPC bar code and have a register "ring up" *unidimensional data* or better yet, *Warning: this data is two-dimensional.* One popular and informal method that can be used however, is to examine a scree plot of the eigenvalues. Although sometimes inconclusive and plagued with issues of nonlinearity and spurious counting of dimensions, the size of the eigenvalues in conjunction with a substantive review of the items can provide insight into how many major traits are being assessed. In any event, practitioners should not let computer programs make decisions for them. An editorial review of test items to characterize or label the dimensions is very essential.

Theory based approaches for assessing dimensionality exist, for example McDonald's non-linear factor analyses (McDonald, 1967) and Holland and Rosenbaum's conditional association approach (Holland and Rosenbaum, 1986). Another approach is that suggested by Nandakumar and Stout's DIMTEST statistical procedure for assessing the lack of *essential unidimensionality* (Stout, 1987; Nandakumar and Stout, 1990). This procedure allows the practitioner to assess whether a specified subtest of items is dimensionally distinct from the remainder of the test. This subtest could be specified either on the basis of content or some exploratory statistical approach (e.g., a principal axis factor analysis). As a first step
this approach can be used to assess whether there is more than one dominant dimension ($d_e > 1$) present in the test. If $d_e > 1$ is indicated, the procedure could be replicated with smaller subsets to help identify the number of dominant dimensions.

After confirming the number of dimensions the multidimensional item parameters can be estimated using a calibration program such as NOHARM II(Fraser, 1983). NOHARM

will only estimate the item parameters however: ability parameters can be subsequently found using Newton-Raphson iterations. One note of caution is necessary: multidimensional item response theory (MIRT) calibration requires a large number of examinees. To obtain satisfactory two-dimensional item parameter estimates it is necessary to have at least 2000 examinees. Such large sample sizes limit the use of multidimensional item response theory to large testing populations such as national programs (e.g., ETS, ACT, ASVAB), statewide testing programs, or large urban school districts.

Once multidimensional item parameters are obtained they can be graphically represented using the work of Reckase (1986). This work provides an excellent foundation for examining the interaction between multidimensional items and the underlying multidimensional ability distributions for groups of interest. For simplicity, in this paper the latent ability space will be taken to be two-dimensional in which one dimension represents the pure, intended-to-be-measured ability, denoted by $\theta$ and the other dimension represents the nuisance abilities, denoted by $\eta$. The ability $\eta$ represents a skill that is **not** intended to be measured, but may be used by examinees to solve an item with a potential for bias.

Reckase's research is based upon the multidimensional item response theory (MIRT) two-parameter logistic (M2PL) model, that for the purposes of this paper, will be expressed in terms of the true ability dimension, $\theta_1$, and the nuisance dimension, $\theta_2$. The probability of a correct response to item *i* by examinee *j* can be written as

$$P(X_{ij}=1 \mid a_i, d_i, \theta_j, \eta_j) = \frac{e^{(a_{1i}\theta_{1j} + a_{2i}\theta_{2j} + d_i)}}{1.0 + e^{(a_{1i}\theta_{1j} + a_{2i}\theta_{2j} + d_i)}} \qquad (1)$$

where $X_{ij}$ is the score (0,1) on item *i* by person *j*, $\mathbf{a}_i$ is the vector of item discrimination parameters, $d_i$ is a scalar difficulty parameter of item *i*, and $(\theta_{1j}, \theta_{2j})$ is the vector of ability parameters for person *j*.

In a two-dimensional latent ability space (e.g., math and verbal ability dimensions), the $a_{1i}$ and $a_{2i}$ vectors designate the composite of $\theta$ and $\eta$ that item *i* is measuring. If $a_{1i} = a_{2i}$, both dimensions would be measured equally well. However, if $a_{1i} = 0$ and $a_{2i} = 1.0$,

discrimination would occur only along the $\theta_2$ dimension with little or no discrimination in the $\theta_1$ direction depending on the correlation between $\theta_1$ and $\theta_2$. If all of the items in a test are measuring exactly the same $(\theta_1, \theta_2)$ composite (i.e., the same "direction" in the $(\theta_1, \theta_2)$ coordinate system), the test would be strictly unidimensional. The more varied the composites that are being assessed, the more multidimensional the test.

Reckase's work describes how to graphically represent an item that requires the application of multiple abilities as vectors in a multidimensional latent space. The length of the vector for item *i* is equal to the degree of multidimensional discrimination, MDISC. This can be computed using the formula

$$MDISC_i = \sqrt{a_{1i}^2 + a_{2i}^2} \tag{2}$$

MDISC is analogous to the unidimensional IRT model's discrimination parameter. The measurement direction of the vector in degrees from the positive $\theta$ axis is

$$\alpha_i = \arccos \frac{a_{1i}}{MDISC_i} \tag{3}$$

This *reference angle* represents the composite of the $\theta_1 - \theta_2$ ability space that item *i* is best measuring. The item vector originates at, and is graphed orthogonal to, the p=.5 equiprobability contour. In the compensatory model described in (1) these equiprobability contours are always parallel.

For item *i*, the distance, $D_i$, from the origin to the p=.5 contour, is computed as

$$D_i = \frac{-d_i}{MDISC} \tag{4}$$

$D_i$ is analogous to the unidimensional IRT difficulty parameter. Because the discrimination parameters are constrained to be positive, the item vectors can lie only in the third quadrant (representing easy items) or in the first quadrant (representing more difficult items). Figure 1 illustrates the item response *surface* for a M2PL item vector whose parameters are: $a_1 = 1.8$, $a_2 = .3$, and $d = .5$. Also illustrated in the bottom portion of Figure 1 is the item's vector, superimposed upon the equiprobability contours of the response surface.

Insert Figure 1 about here

Once the vectors for an item are graphically displayed the validity sector can be defined. Qualitatively, it is simply the sector which contains the most homogeneous subset of item vectors. How wide the sector should be, or which items should be included, relies upon the judgement of the testing practitioner. Depending on the specificity of the trait being measured the width of the validity sector will vary. The better a skill can be defined the narrower the sector; the more ambiguous the trait the wider the sector. Ongoing research is studying the relationship between Nandakumar and Stout's test of essential unidimensionality, item discrimination, and sector width.

Detailed below are two explicit examples of how MIRT analyses and the concept of the validity sector can be used in testing research. The intention is not to contrive one more statistical analysis to add to the numerous analyses that researchers can perform on test data, but rather, to develop a coherent methodology that will help provide a common basis of discussion between item writers and measurement specialists. The goal is to provide a mechanism by which the content and construct specifications of a test can be statistically verified.

*Using the validity sector to confirm content differences*

One major use of the validity sector would be to provide a statistical validation of the table of content specifications used by the item writers. For illustration purposes, a subset of items from a 60-item ACT Assessment Mathematics Usage Test was calibrated using NOHARM to fit the M2PL model. This subset of items was created according to a two-way table of specifications. There were three levels of content: Pre-Algebra, Elementary Algebra, and Intermediate Algebra, and two levels of skill: Basic Skills and Application.

Pre-Algebra items included items which required the examinee to algebraic operations with whole numbers, decimals, and fractions. Elementary Algebra items involved operations with algebraic expressions, including the factoring of quadratic equations. The third category, Intermediate Algebra items were based on operations using integer exponents, radical expressions, linear inequalities, and solving systems of two linear equations. The Preliminary

Technical Manual for the Enhanced ACT Assessment (1989) states that the Basic Skills items were designed to be solved by "performing a familiar sequence of operations in a familiar setting" (p.17). Solutions to Application items are expected to be obtained by executing a "familiar sequence of operations, but the solution [would] not be routine" (p.17).

Item vector plots (obtained using NOHARM II) for each six cells of the table of specifications are shown in Figure 2. Based upon these plots it appears that the Pre-Algebra/ Basic Skills items are the most homogeneous and contain items whose vectors lie in a very narrow sector ( $17°$ ). Except for the number of items, the Elementary and Intermediate Algebra items do not appear to have any distinguishing characteristics. Both contain items that have a wide range of measurement composites. For each of these contents, the Application items are more discriminating (i.e., have longer vectors) than the Basic Skills items.

---

Insert Figure 2 about here

---

A review of the actual items, (insight which could be reinforced by item writers), suggests that items with large measurement angles tend to be more equation-oriented and contain little text. A typical item of this type would be:

Which of the following expresses 60 as a product of prime numbers?

a. $2 \times 3 \times 5$
b. $2 \times 2 \times 15$
c. $2 \times 2 \times 3 \times 5$
d. $2 \times 3 \times 3 \times 5$
e. $1 \times 2 \times 5 \times 6$

Note the above item had a measurement angle of $82°$ . Based upon the types of items that were measuring the $\theta_2$-axis the second dimension was labeled as an algebraic symbol manipulation skill. It should also be noted that the easier items, those which lie in the third quadrant, tend to discriminate primarily between levels of this ability.

Elementary and Intermediate algebra items which had very small measurement angles contained much longer item stems and could be classified as "story problems".

An example of an item which measured mostly $\theta_1$ is

> Joe has taken 4 tests in his algebra class during the current grading period, earning test scores of 86, 66, 78, and 81. A student needs an average score of 80 on 5 tests to earn a "B" for the class. What is the minimum (integer) score Joe can earn on his next test in order to have an average of at least 80 for the 5 tests?
>
> a. 83
> b. 85
> c. 87
> d. 89
> e. 91

This item had a measurement angle of $12°$. Most of the items with low measurement angles required the examinee to read a verbal description, translate the problem into a algebraic form, and solve. Hence, the $\theta_1$ axis was labeled as a verbal skill and algebraic translation dimension. A review of the vector plots indicates that this dimension appears to be measured by the more difficult items. One might wonder if there is a confounding of difficulty and dimensionality (cf. Davey, Ackerman, & Reckase, 1989).

The information presented in Figure 1 could also be quantified. An example is displayed in Table 1. This table lists the means and standard deviations of the MDISC and difficulty values for the items in each cell. Also provided is the range of the measurement angles for each cell, and the width of the sector that encompasses all of the items. This table seems to confirm the vector plots shown in Figure 2. The Intermediate Algebra items have the widest sector, $58.42°$, and Elementary Algebra items the narrowest, $45.29°$. The Application Elementary Algebra items were the most discriminating items, Basic Skills Elementary Algebra items were the least discriminating, $\mu_{MDISC} = 1.63$ and .94 respectively. Pre-Algebra items were, on average, the easiest set of items, $\mu_{d_i} = .15$, Elementary Algebra items the most difficult, $\mu_{d_i} = -.72$. Basic Skills items tended to be easier and less discriminating than the Application items.

Insert Table 1 about here

Conceptually, one might think that each cell of the table should be measuring a unique combination of skills in the latent space. Likewise, plots of the content classifications (collapsed over skill level) should contain items with unique, non-overlapping sectors. However, based upon the item vector plots this does not appear to be the case. One possible answer to explain what seems to be happening is that the items are not distinguishing between levels of "Pre-Algebra-ness" or "Elementary Algebra-ness" but rather some higher level skill that the cuts across the content categories. For example, the skill to manipulate algebraic equations is not a content category but rather a proficiency that transcends all of the algebra categories. The same argument might apply to the skill levels when collapsed over content. The item vectors for each content and each skill level are shown in Figures 3 and 4.

Insert Figures 3 and 4 about here

As in the item vector plots for each content, there does not seem to be a unique sector of the ability plane that is defined by each skill. These plots appear to raise more questions about what is actually being measured than they answer. If an item writer were asked what the test was measuring, the reply would probably be related to the content specifications from which the items were created. If a psychometrician were asked the same question, the answer, if based upon item vector plots, might detail higher order skills that run across the various contents. So what is the real name of this test anyway? One could play it safe and call it a general mathematics usage test but that wouldn't provide the examinee, nor the colleges which want to use the scores, with much insight. But then again, maybe the vector plots are only providing a resurrection of what the original test developers from ACT discovered long ago and hence decided to call it a "mathematics usage" test. But hopefully, our analyses have become more sophisticated an enable us to be more specific about what it is we are measuring and what it is we are not measuring.

The discussion above is intended to serve as a staring point of dialogue between the item

writer and the psychometrician. There are many questions and concerns that need to be raised: Most importantly, do we have the right dimensionality? (Hirsch and Reckase (1991) provided excellent examples of how the orientation of item vectors can be totally reversed if the modeled dimensionality is less than the real dimensionality.) What is the relationship between item characteristics (i.e., wording, length of text, use of figures, content topic) and the measurement angle? Is it possible to redefine the specifications of particular cells to make the measurement composites of the items more homogeneous? Is it possible to construct easy items which measure mostly the second ability dimension (i.e., algebraic translation of text)? The mutual goal should be to identify the characteristics that produce items measuring certain ability composites. It is expected that such a process will be iterative and require the experimentation of many hypothesized relationships. The ultimate goal would be to have the psychometrician provide a statistical affirmation of what the item writers believe their items are measuring.

*Using the validity sector to detecting biased items*

To examine bias within a multidimensional framework, the true and the nuisance ability dimensions need to be identified and handled separately. The true ability can be thought of as some hypothetical theoretical ability (or linear composite of multiple abilities) that a test is designed to measure. In reality, no matter how carefully test items are written, they have a propensity to measure nuisance abilities as well. Nuisance abilities can be thought of as skills or content information which the examinee needs to solve particular items but were not intended to be assessed by the item writer. For example reading ability may be considered to be a nuisance skill in a test designed to measure the pure ability of algebraic symbol manipulation. Shealy and Stout (1991) proposed that researchers consider the conditional distribution of the nuisance ability for each level of the valid ability. If this conditional distribution of the nuisance ability differs across groups of interest, the potential for bias exists.

The issue of item bias and construct validity are interrelated. That is, the number of skills being measured and the degree to which comparisons between groups are appropriate is a construct validity issue. If a test lacks construct validity, then it is quite likely that some of the items are measuring supplementary skills and the interaction between examinees and the examinee and these items could result in bias. Bias will be realized if groups of interest differ in their

underlying conditional distributions of these extraneous skills. Simply put, having items on a test that are construct invalid is a necessary, but not sufficient cause of item bias.

If all the items are measuring only the valid skill or construct, and the item-examinee interaction is unidimensional, then any group ability differences will be due to *impact*, not bias. Impact can be formally defined as a between-group difference in test performance caused by group ability differences on the valid skill (e.g., the differences between the proportion correct for two groups of interest on a valid item).

Shealy and Stout (1991) suggested that in performing a bias analysis the practitioner identify the most valid items for creating a conditioning test score to match examinees. For each test there should be a specified sector which envelops only those item vectors that are measuring the composite of abilities the test was designed to measure for both groups. That is, all items measuring these composites in this user specified sector may be considered to be valid. Figure 5 displays an item vector plot from a subset of items from a ACT Mathematics Usage Test (Reckase, 1985). Outlined in this plot is a suggested validity sector. Notice that items that lie outside of this sector discriminate mostly between levels of $\theta_2$ proficiency and therefore have the potential for eliciting bias. This bias would be manifested if the two groups of interest differed on this secondary skill. Consequently, these items, referred to as construct invalid items, would be the suspect items in an item bias study. Once identified the practitioner could use either the Mantel-Haenszel procedure (Holland & Thayer, 1990) or the Simultaneous Item Bias (SIB) detection procedure developed by Shealy and Stout (1991) if testing the items one at a time seems appropriate. It should be noted that SIB also does have the capability to test for bias in multiple items simultaneously.

---

Insert Figure 4 about here

---

Ackerman (1992) has demonstrated that by using the items that lie within a narrow validity sector the reference composites (Wang, 1986) for the two groups of interest in a bias analysis will be quite similar. The reference composite is a linear approximation to the direction of measurement in the latent space being measured by the unidimensional based maximum likelihood estimate. This is paramount to any bias study. If the reference composites for the two

groups lie in different directions the interpretation of the score scale for each group would have quite different meanings and any subsequent bias analysis that matched subjects would be comparing "apples with oranges".

## Conclusion

The purpose of this paper was to provide insight into how the item writer and the psychometrician could use multidimensional item response theory to evaluate the construct validity of a test. Two examples were provided in which the information provided by validity sectors was used to improve the understanding of the measurement process. In the first instance, the table of specifications of a mathematics usage test was explored by examining the item vectors for each cell in the table. By examining the width and degree of overlap of the validity sectors one can gain insight about the amount of unique and redundant information different content or skill levels are providing. Ideally, psychometricians working with item writers can eventually establish the characteristics of items which cause them to measure particular latent ability composites.

The second example demonstrated how items with the greatest potential for eliciting bias can be easily determined by constructing a plot of the item vectors and establishing a validity sector. Such an analysis should be coordinated with subsequent MH or SIB bias analyses.

The intent of this paper is **not** to suggest changes in the job descriptions of either item writers or psychometricians. It is expected that the psychometrician, based upon the level of training and expertise would conduct the MIRT analysis and create the item vector plots. But, it is hoped that item writers (in concert with psychometricians) would be able to understand what the item vector plots mean and hopefully provide a substantive interpretation.

One should never overlook the difficulty of multidimensional analyses and the problem of determining the correct number of interpretable dimensions. In this paper all of the multidimensional examples assumed there were only two dimensions. If there are three identifiable dimensions, the validity sector would become a "validity cone". In more than three we might consider a "hyper cone".

The intent of the validity sector concept is to help bridge the conceptual and statistical gap between item writers and psychometricians. The closer the statistical analyses can come to verifying the underlying test specifications the more accurately one can explain to an examinee what skills are being assessed and how accurately. Some day it would be nice if the story an item writer would tell an examinee about what a test measured was the same story a psychometrician, talking about the same test and in the same language, would tell.

## References

Ackerman, T. A. (1992). An explanation of differential item functioning from a multidimensional perspective. Journal of Educational Measurement 24, 67-91.

American College Testing Program. (1989). Preliminary Technical Manual for the Enhanced ACT Assessment. Iowa City, IA: The American College Testing Program.

Davey, T.D., Ackerman, T.A. & Reckase, M.D. (1989). The interpretation of score differences when item difficulty and dimensionality are confounded. Paper presented a the Annual Meeting of the Psychometric Society, Los Angeles, CA.

Fraser, C. (1983). NOHARM II: A FORTRAN program for fitting unidimensional and multidimensional normal ogive models of latent trait theory. Amidale, Australia: University of New England, Centre for Behavioral Studies.

Holland, P. W. & Rosenbaum, P. (1986). Conditional association and multidimensionality in monotone latent variable models. Annuals of Statistics, 14, 1523-1543.

Holland, P.W. & Thayer, D.T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H.I. Braun (Eds.), Test Validity (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hirsch, T. & Reckase, M. D. (1991). Projections of higher dimensional MIRT solutions on lower dimensional spaces. Paper presented at the Office of Naval Research Contractor's Meeting,Princeton, NJ.

McDonald, R. (1967). Nonlinear factor analysis. Psychometric Monograph, 15.

Nandakumar, R. & Stout, W. (in press). Refinements of Stout's procedure for assessing latent trait dimensionality. Journal of Educational Statistics.

Reckase, M.D. (1985, April). The difficulty of test items that measure more than one ability. Applied Psychological Measurement, 9, 401-412.

Reckase, M.D. (1986, April). The discriminating power of items that measure more than one dimension. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Shealy, R. & Stout, W. (in press). An item response theory model for test bias. (ONR Technical Report); In H. Wainer & P. Holland (Eds.), Differential Item Functioning, Theory and Practice, Hillsdale, NJ: L. Erlbaum Associates.

Shealy, R. & Stout, W. (1991). A procedure to detect test bias present simultaneously in several items. (Technical Report 91-3-ONR). Champaign, IL: University of Illinois.

Stout, W. (1987). A nonparametric approach for assessing latent trait undimensionality. Psychometrika, 52, 589-617.

Wang, M. (1986). Fitting a unidimensional model to multidimensional item response data. Paper presented at the annual meeting of the Office of Naval Research Contractors, Knoxville, TN.

Table 1
Mean and standard deviation MDISC and difficulty values, measurement range and sector width for the items in each cell of the table of specifications.

| Content | | Skill: Basic Skills | Skill: Application | Total |
|---|---|---|---|---|
| Pre-algebra | $\mu_{MDISC}$ | .96 (.26)[a] | 1.13 (.43) | 1.00 (.39) |
| | $\mu_{d_i}$ | .64 (.66) | -.20 (.91) | .15 (.91) |
| | α range[b] | 66.50 - 83.50 | 36.52 - 85.29 | 36.52 - 85.29 |
| | Sector Width[b] | 17.00 | 48.77 | 48.77 |
| | n[c] | 4 | 8 | 12 |
| Elementary algebra | $\mu_{MDISC}$ | .94 (.40) | 1.63 (.36) | 1.25 (.52) |
| | $\mu_{d_i}$ | -.19 (.75) | -1.40 (1.57) | -.72 (1.33) |
| | α range | 32.76 - 67.09 | 21.80 - 57.81 | 21.80 - 67.09 |
| | Sector Width | 34.33 | 36.01 | 45.29 |
| | n | 5 | 5 | 9 |
| Intermediate algebra | $\mu_{MDISC}$ | 1.06 (.44) | 1.34 (.73) | 1.23 (.64) |
| | $\mu_{d_i}$ | -.47 (1.28) | -1.10 (1.16) | -.08 (1.42) |
| | α range | 25.59 - 75.41 | 18.74 - 77.16 | 18.74 - 77.16 |
| | Sector Width | 49.82 | 58.42 | 58.42 |
| | n | 7 | 9 | 16 |
| Total | $\mu_{MDISC}$ | 1.00 (.39) | 1.33 (.59) | |
| | $\mu_{d_i}$ | -.11 (1.09) | -.82 (1.39) | |
| | α range | 25.59 - 83.50 | 18.74 - 85.29 | |
| | Sector Width | 57.90 | 66.55 | |
| | n | 16 | 21 | |

Note: [a]denotes standard deviation value
[b]values are in degrees
[c]number of items

## Figure Captions

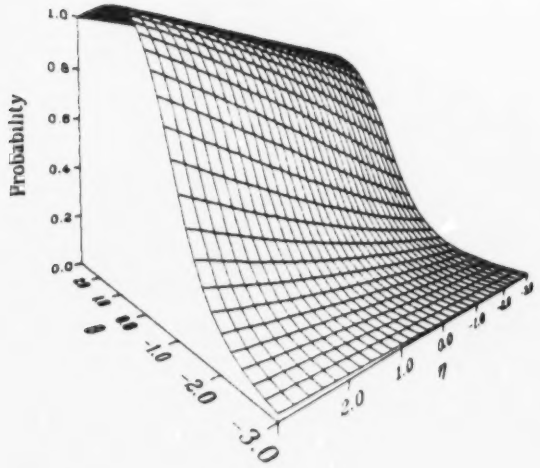
Figure 1. The item response surface and corresponding contour with the item vector for the M2PL parameters, $a_1 = 1.8$, $a_2 = .3$, $d = .5$.

Figure 2. Item vectors for each cell in the two-way classification scheme of the table of specifications.

Figure 3. Item vectors for each of the three content categories.

Figure 4. Item vectors for each of the two skill categories.

Figure 5. A validity sector detailing construct valid items and potentially biased items.


Probability
1.0
0.8
0.6
0.4
0.2
0.0
θ
-1.0
-2.0
-3.0
1.0
2.0
η

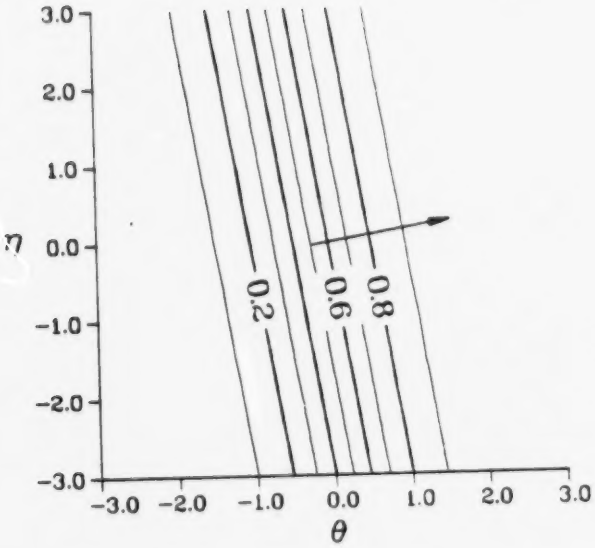

3.0
2.0
1.0
η 0.0
-1.0
-2.0
-3.0
-3.0 -2.0 -1.0 0.0 1.0 2.0 3.0
θ
0.2
0.6
0.8

Figure 1

Figure 2

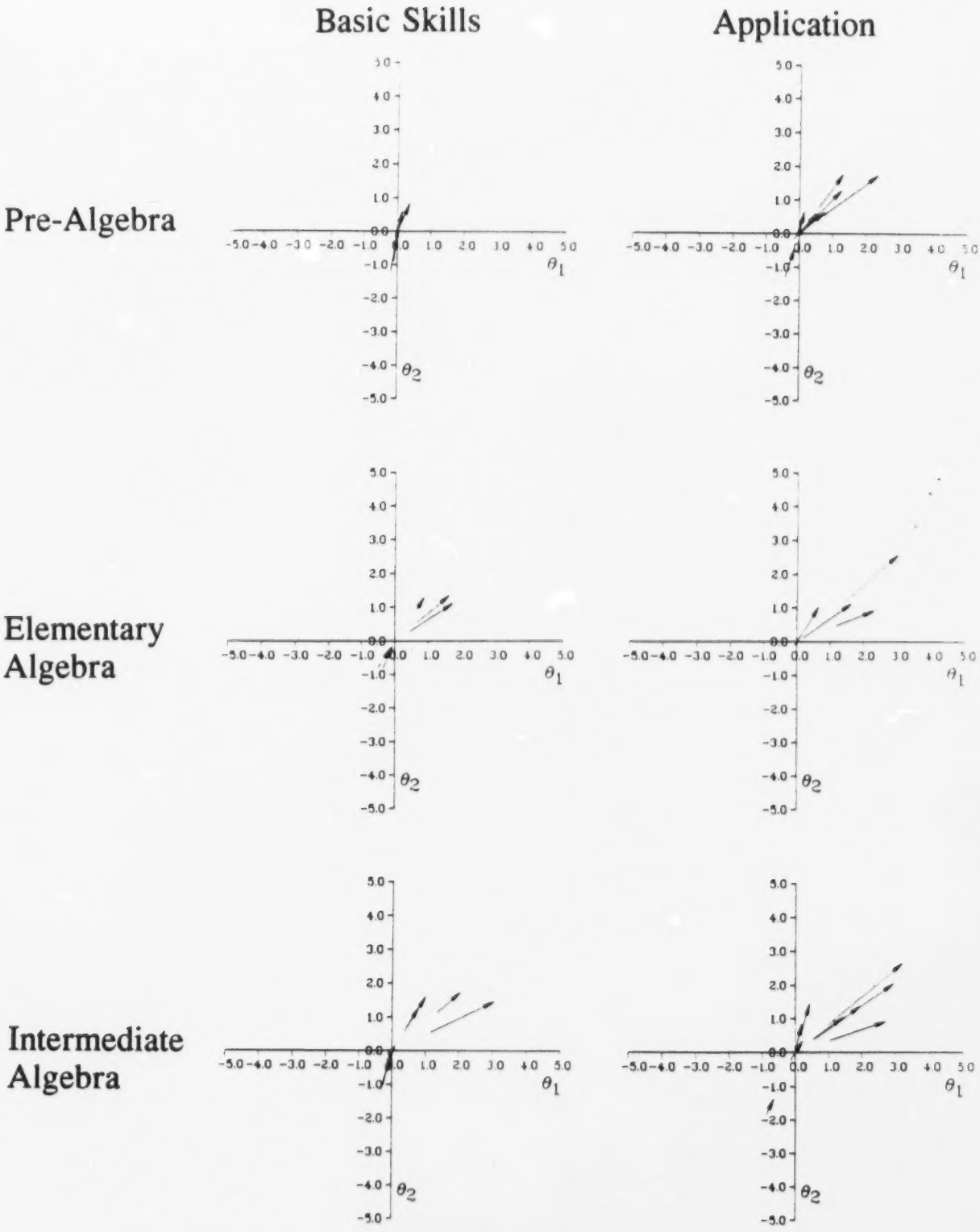
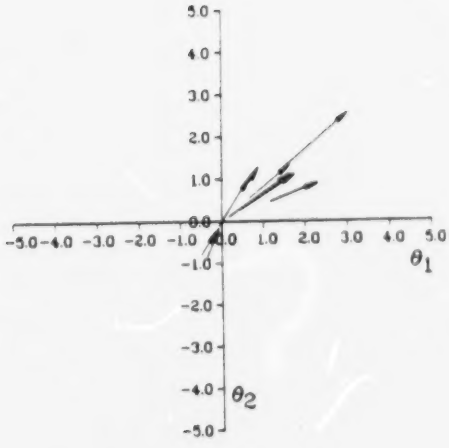

Basic Skills
Application
Pre-Algebra
Elementary Algebra
Intermediate Algebra
θ1
θ2

Figure 3

Pre-Algebra

Elementary Algebra




Intermediate Algebra

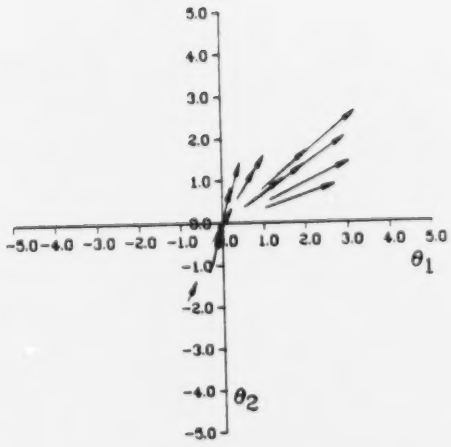
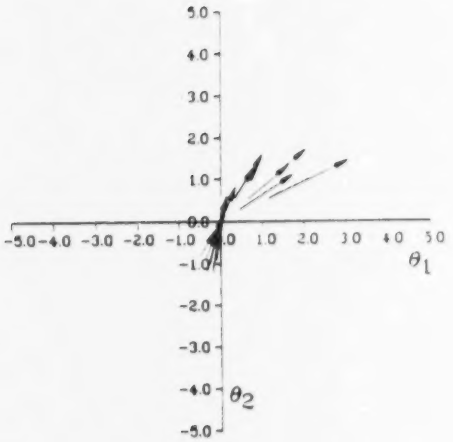
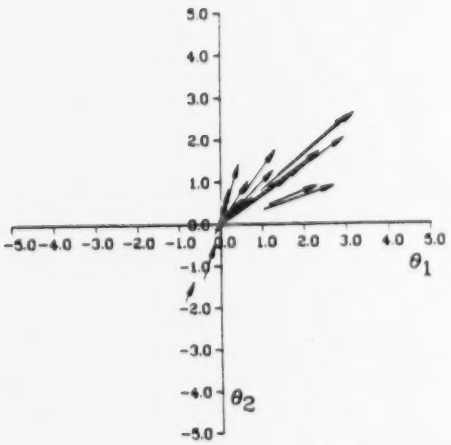


Figure 4

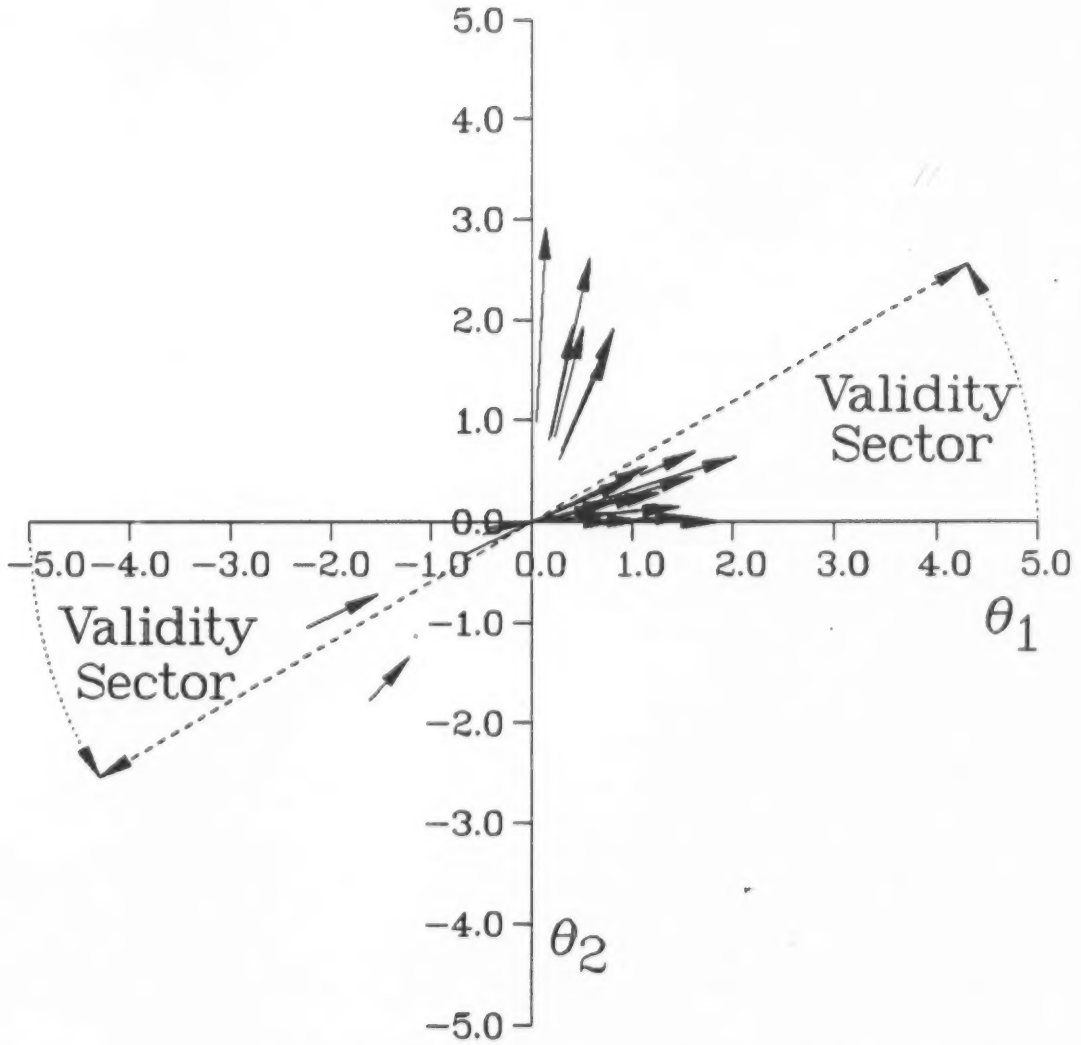



Basic Skills




Application

Figure 5


5.0
4.0
3.0
2.0
1.0
0.0
−1.0
−2.0
−3.0
−4.0
−5.0
−5.0 −4.0 −3.0 −2.0 −1.0 0.0 1.0 2.0 3.0 4.0 5.0
Validity Sector
Validity Sector
$\theta_1$
$\theta_2$